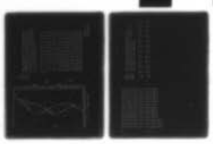
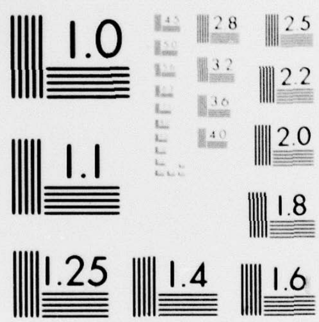
AD-A073 269 TEXAS A AND M UNIV COLLEGE STATION INST OF STATISTICS F/G 12/1
DISCUSSION OF PAPER BY GOOD AND GASKINS ON DENSITY ESTIMATION A--ETC(U)
JUL 79 E PARZEN DAAG29-78-G-0180
UNCLASSIFIED TR-A-11 ARO-16228.11-M NL

| OF |

AD
A073269



END
DATE
FILMED
9-79
DDC



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

TEXAS A&M UNIVERSITY
COLLEGE STATION, TEXAS 77843

6

Discussion of Paper by Good and Gaskins
on Density Estimation and Bump Hunting,

10

by Emanuel Parzen
Institute of Statistics, Texas A&M University

9 14 TR–
Technical Report, No. A-11

11 July 1979

12 14p.

12

Texas A & M Research Foundation
Project No. 3861
"Maximum Robust Likelihood Estimation and
Non-parametric Statistical Data Modeling"
Sponsored by the U.S. Army Research Office

Professor Emanuel Parzen, Principal Investigator

Approved for public release; distribution unlimited.

ARO 16228.11-M

Unclassified
SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER Technical Report No. A-11	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) Discussion of Paper by Good and Gaskins on Density Estimation and Bump Hunting		5. TYPE OF REPORT & PERIOD COVERED Technical
		6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s)		8. CONTRACT OR GRANT NUMBER(s) 15 DAAG29-78-G-0180
9. PERFORMING ORGANIZATION NAME AND ADDRESS Texas A&M University Institute of Statistics College Station, TX 77843		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS Army Research Office Research Triangle Park, NC 27709		12. REPORT DATE July 1979
		13. NUMBER OF PAGES 26
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office) 18 ARO 19 16228.11-M		15. SECURITY CLASS. (of this report) Unclassified
		15a. DECLASSIFICATION DOWNGRADING SCHEDULE
16. DISTRIBUTION STATEMENT (of this Report) Approved for public release; distribution unlimited.		
17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report) NA		
18. SUPPLEMENTARY NOTES The findings in this report are not to be construed as an official Department of the Army position, unless so designated by other authorized documents.		
19. KEY WORDS (Continue on reverse side if necessary and identify by block number) Density estimation Density-quantile functions Statistical data modelling		
20. ABSTRACT (Continue on reverse side if necessary and identify by block number) The view is expressed that the best way to evaluate the approach of Good and Gaskins is to compare it with alternative approaches to density estimation currently available. As a contribution towards this comparative study, this paper describes an analysis of their data using the density-quantile estimation approach proposed by Parzen (1979).		

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE
S/N 0102-LF-014-6601

Unclassified
SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

Discussion of I. J. Good and R. A. Gaskins

"Density Estimation and Bump Hunting by the Penalized Likelihood Method Exemplified by Scattering and Meteorite Data"

by

Emanuel Parzen *

Texas A&M University

It gives me great pleasure to discuss a paper on the estimation of probability density functions and the location of bumps. There is an extensive literature on density estimation but many statisticians seem doubtful about the usefulness of these techniques because their application seems subjective and complicated. A major criticism I would make of this paper is that it does not help to dispel this negative attitude of statisticians towards density estimation. One cannot help but be impressed by the ingenuity of Drs. Good and Gaskins, and even to believe that they may be able to successfully fit probability densities to data. But I have doubts if other statisticians would find their approach a practical method for daily use in statistical data analysis.

I find it strange that the authors would dismiss the "cross-validation" method of Grace Wahba on the ground that it requires statisticians "to understand much more advanced mathematical concepts than we need in the present work." I believe that statistical computing has changed the statisticians' criterion for what makes a statistical technique effective; it is not how easy it is to compute by hand, or how easy it is to understand in a rigorous sense the theory of the technique, but how easy it is for a statistician to communicate to a client how to interpret the computer output (hopefully graphical) of a program implementing the techniques. I conjecture that Good and Gaskins may have more work to do to convince statisticians of the practical effectiveness of their techniques of density estimation. The unique aspect of the theory of G and G' is the evaluation of the probabilities of bumps; I do not believe one can easily understand either the philosophy or the methods they propose for this problem.

Many choices need to be made to apply the G and G' theory: a

* Emanuel Parzen is Distinguished Professor at the Institute of Statistics, Texas A&M University, College Station, TX 77843. This research was supported in part by the Army Research Office, Grant DAAG-29-78-C-0180.

smoothing parameter B, a Fourier series expansion truncation point r, and a linear scale transformation of the data. Truncation points as large as 271 and 601 are surprisingly large; perhaps without justification, statisticians wonder if one can't fit anything one wants with so many parameters.

I believe that the best way to evaluate the approach of G and G' is to compare it with alternative approaches to density estimation currently available. As a contribution towards this comparative study, I should like to describe an analysis of their chondrite and LRL data using the density-quantile estimation approach proposed in Parzen (1979).

The density-quantile estimation approach applies equally well to ungrouped or grouped data. Step I is to form a raw estimator $\tilde{Q}(u) = \tilde{F}^{-1}(u)$ of the quantile function, or inverse distribution function, $Q(u) = F^{-1}(u)$ of the data. Step II is to form a raw estimator $fQ(u)$ of the density-quantile function $fQ(u) = f(F^{-1}(u)) = 1/Q'(u)$ by choosing a grid value h (=.01, say) and define

$$\tilde{f}Q(jh) = \frac{2h}{\tilde{Q}((j+1)h) - \tilde{Q}((j-1)h)}, \quad j=1, 2, \ldots, H-1$$

where $H = 1/h$ is assumed to be an integer. Step III is to form successive smooth estimators $\hat{f}Q_m(u)$, $m=1, 2, \ldots$, by forming successive autoregressive smoothers $d_m(u)$ of

$$\tilde{d}(u) = \frac{f_0Q_0(u)\ \tilde{Q}'(u)}{\int_0^1 f_0Q_0(t)\ \tilde{Q}'(t)dt}$$

where $f_0Q_0(u)$ is a suitable base density-quantile usually chosen equal

to $\phi\Phi^{-1}(u)$ the standard normal density-quantile function. Steps I-III can be implemented routinely and require no choices by the statistician. The crucial question of how to choose the order m could be regarded as an open question; it could be based on a graphical comparison of how well $\tilde{D}(u) = \int_0^u \tilde{d}(t)dt$, $0 \le u \le 1$ is fitted by $\hat{D}_m(u) = \int_0^u \hat{d}_m(t)\,dt$, or on the decay to zero of the Fourier transforms or pseudo-correlations

$$\tilde{\rho}(v) = \int_0^1 e^{2\pi iuv}\ \tilde{d}(u)\,du, \quad v = 0, \pm 1, \ldots .$$

For the chondrite data, the shape of $\hat{f}Q_m(u)$ is the same for $2 \le m \le 5$, and the fit of $\hat{D}_m(u)$ to $\tilde{D}(u)$ is close for all these values of m. The density is trimodal. From the graph of the density-quantile function, one can determine the percentiles p_j, $j=1, 2, 3$, at which bumps occur; the x values at which these bumps are located are determined from the sample quantile function by $x_j = \tilde{Q}(p_j)$, $j=1, 2, 3$.

For the LRL data, the fit of $\hat{D}_m(u)$ of $\tilde{D}(u)$ is the same for, say $5 \le m \le 30$ (the largest order we examined). We believe it desirable to consider the estimators $\hat{f}Q_m(u)$ corresponding to two values of m; possible criteria for interesting values of m are: the smallest value for which $\hat{D}_m(u)$ well fits $\tilde{D}(u)$, and the smallest value of m beyond which $\tilde{\rho}(v)$ is not "approximately" zero. Table I of values of $|\tilde{\rho}(v)|^2$ leads us to consider $m=8$ and $m=15$. From the graphs of $\hat{f}Q_8(u)$ and $\hat{f}Q_{15}(u)$ one could identify bumps and inflection points corresponding to groups (iii) - (ix) claimed by G and G'. However one must adapt the density-quantile technique to see evidence of groups (i), (ii), and (x) - (xiii), because these occupy respectively the bottom 2% and top 2% of the data.

The existence of bumps at the extremes of a sample can be investigated for large sample sizes by treating the bottom and top ends of the original sample as two new samples to be analyzed by themselves. As the bottom end sample, we take the data up to 550, which includes group (iii) in (525,535); the top end sample we consider consists of the data starting with 1555, which includes group (ix) in (1585, 1625). The order of the density-quantile estimator was chosen equal to 3 in both cases on the basis of the decay of their pseudo-correlations $\rho(\nu)$; graphs of these estimators are shown in Figures L and M, and both indicate 4 bumps. The raw density-quantile function is very wiggly for these two samples and the estimated density-quantile function exhibit many bumps as the order is increased. Thus it is important to have further research on the problem of determining the "best order" or amount of smoothing.

To justify the technique of analyzing the ends of a sample by itself note that we are analyzing $Q(u)$, $0 \le u \le u_1$, and $u_2 \le u \le 1$ for specified percentiles u_1 and u_2 by analyzing the quantile functions Q_1 and Q_2 on $0 \le u \le 1$ defined by

$$Q_1(u) = Q(uu_1) , \quad Q_2(u) = Q(u_2+(1-u_2)u)$$

with derivatives $q_1(u) = u_1 q(uu_1)$, $q_2(u) = (1-u_2)q(u_2+(1-u_2)u)$;
consequently

$$f_1Q_1(u) = \frac{1}{u_1} fQ(uu_1), \; f_2Q_2(u) = \frac{1}{1-u_2} fQ(u_2+(1-u_2)u) .$$

These formulas agree with the amplitudes indicated on the graphs. One thus can rescale $f_jQ_j(u)$ to make it an extension of $fQ(u)$ at the ends of the interval.

The foregoing discussion of estimators of the density-quantile functions of the data analyzed by G and G' has emphasized how our results tend to confirm the results of G and G', using techniques whose use seems more able to be routinized. Next we would like to point out the possibility of an important bump not found by G and G'. If one chooses $m \ge 6$, the density-quantile estimator of all the LRL data finds an additional bump not disclosed by G and G'; it appears that in the region of their group (iv) the density-quantile function (as well as $\hat{f}Q$) is bimodal, and there are two distinct groups (splitting at 775) within group (iv). It should be noted that if one chooses order $m = 7$ or less, one would see only one mode in group (iv). From $\hat{f}Q_7(u)$ one infers the presence of fewer bumps. I believe the statistician should present $\hat{f}Q_m(u)$ to the physicist for various values of m, and collaborate with him using scientific theoretical considerations, as well as statistical tests, to determine which smoothings better fit the facts.

Statistical scientists should heed the flippant advice of Sir Arthur Eddington: "Never trust an experimental result until it has been confirmed by theory". Nevertheless I believe that we can develop confident means of statistical model identification which start with a body of collected data and seeks to learn by analyzing the data under changing models (a computer environment version of observing a phenomenon under changing conditions in order to learn what happens when a factor is varied in a controlled way).

The existence of bumps should thus be checked by estimating the density-quantile function of the sample quantile function on a sub-interval. The LRL data from 565-1115 (which includes bumps (iv), (v), (vi) found by

G and G') was analyzed; its $\hat{f}Q_3(u)$, shown in Figure N, indicates only the bumps found by G and G'.

Thus the density-quantile estimation approach would agree with the conclusions drawn by G and G' on the chondrite and LRL data. The main question I would like to ask is whether their technique can be packaged for widespread reliable use and, most importantly, communication of insight.

Reference

Parzen, E. (1979) "Nonparametric statistical data modeling", Journal of the American Statistical Association, 74, 105-131.

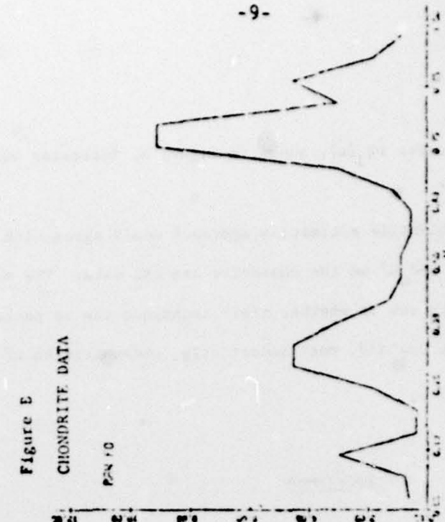
Table 1

LRL Data

Squared Modulus $|\tilde{\rho}(v)|^2$ Used to Select Orders 8 and 15

v	$\|\rho(v)\|^2$	v	$\|\rho(v)\|^2$
0	1.0000	16	.0000
1	.1268	17	.0002
2	.0076	18	.0002
3	.0137	19	.0001
4	.0158	20	.0002
5	.0113	21	.0001
6	.0016	22	.0006
7	.0013	23	.0004
8	.0045	24	.0006
9	.0019	25	.0001
10	.0006	26	.0002
11	.0021	27	.0003
12	.0012	28	.0001
13	.0000	29	.0001
14	.0005	30	.0000
15	.0006		

List of Figures



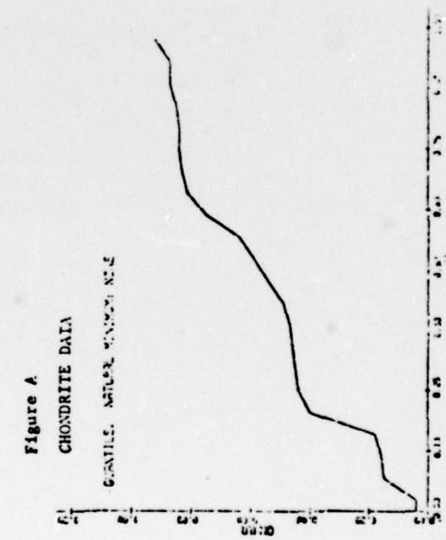


Figure C

CHONDRITE DATA

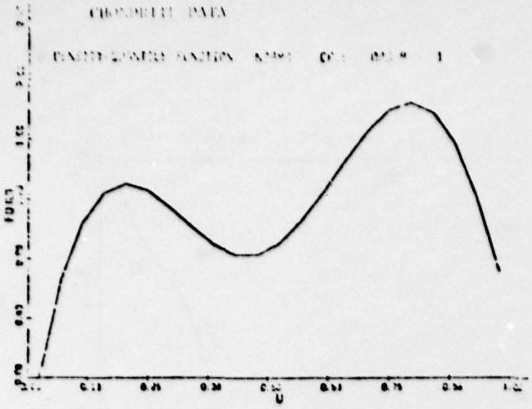


Figure E

CHONDRITE DATA

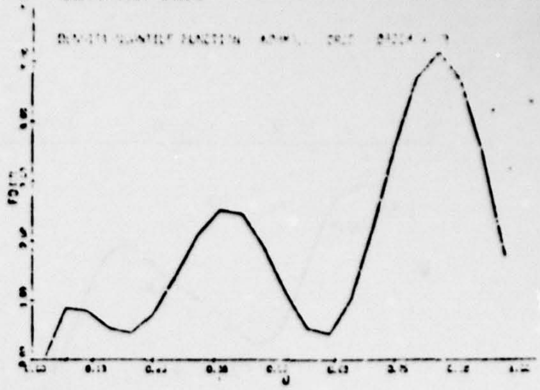


Figure D

CHONDRITE DATA

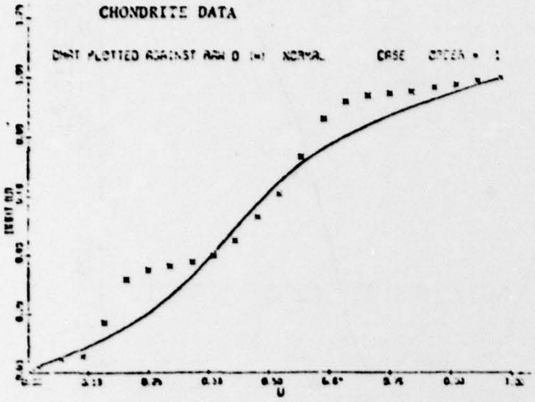


Figure F

CHONDRITE DATA

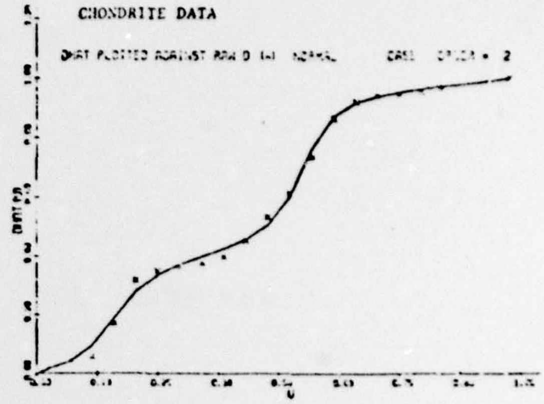
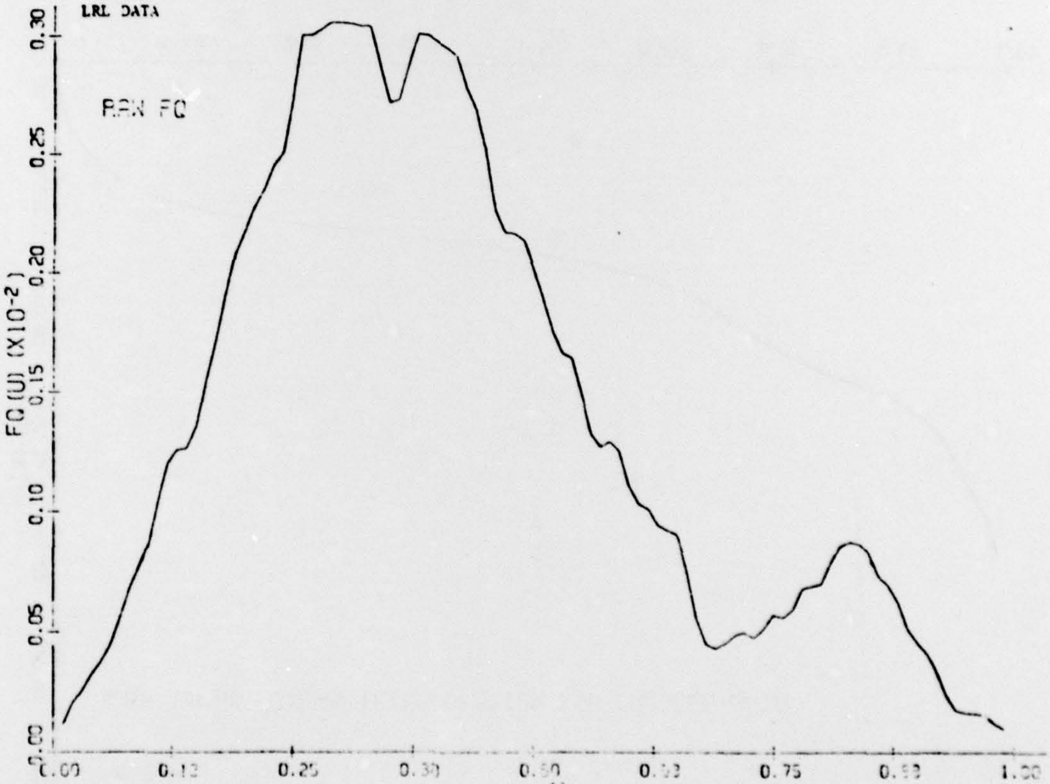


Figure G

LRL DATA



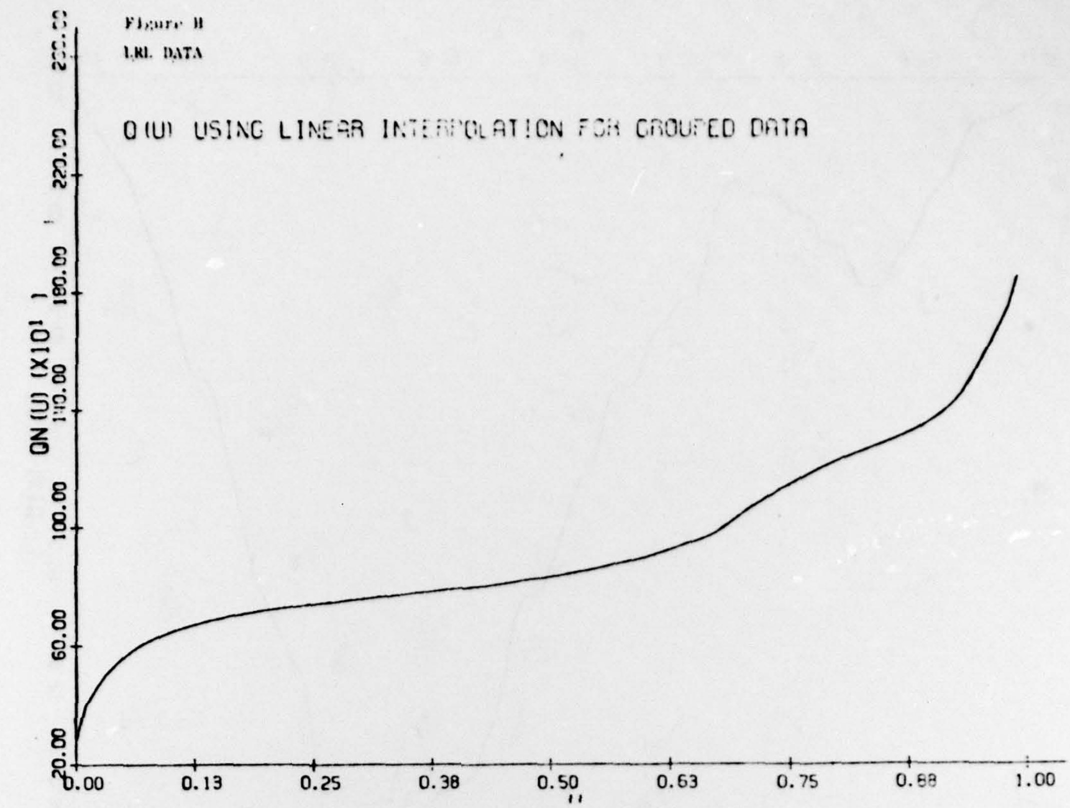


Figure II
LRL DATA

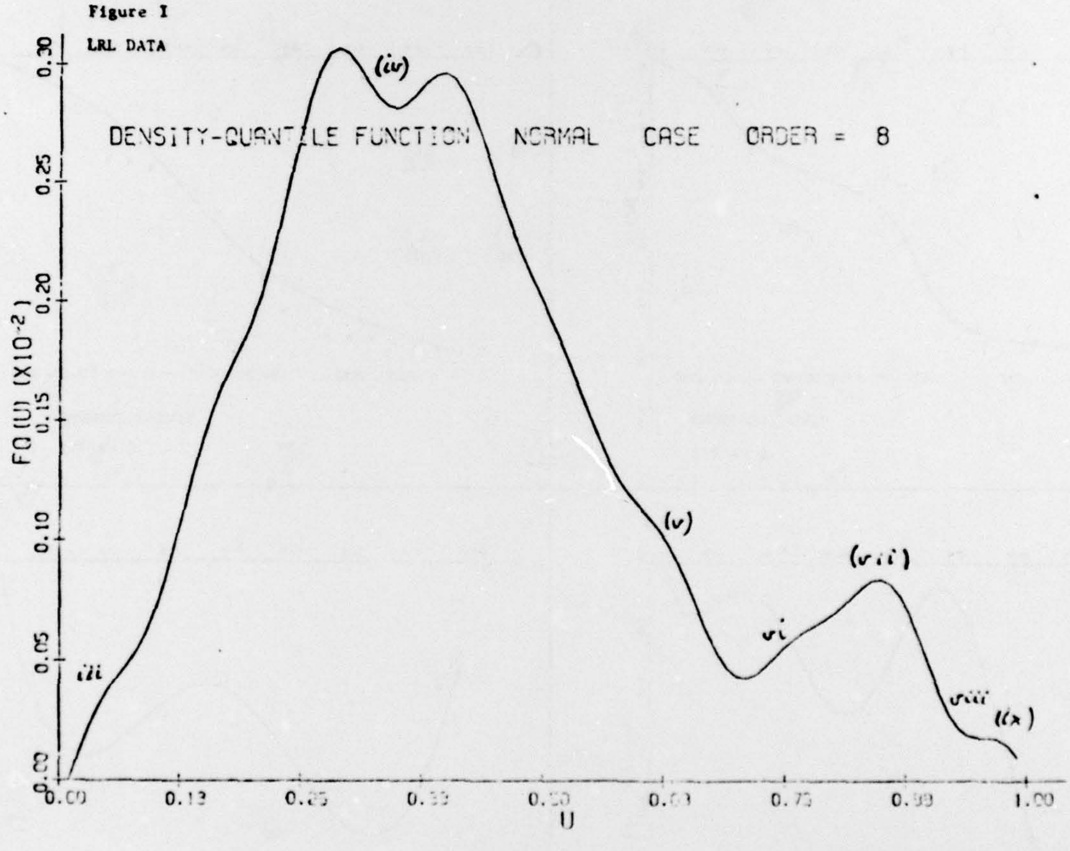


Figure I
LRL DATA

Figure J

LRL DATA

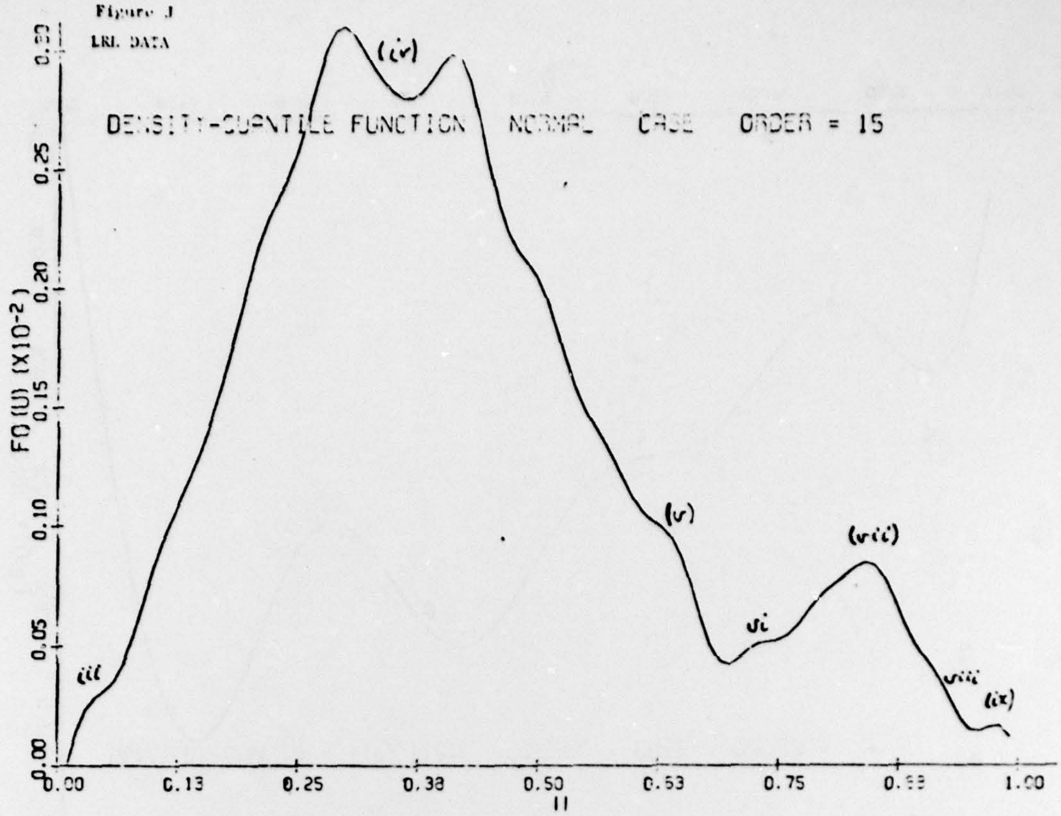


Figure K

LRL DATA

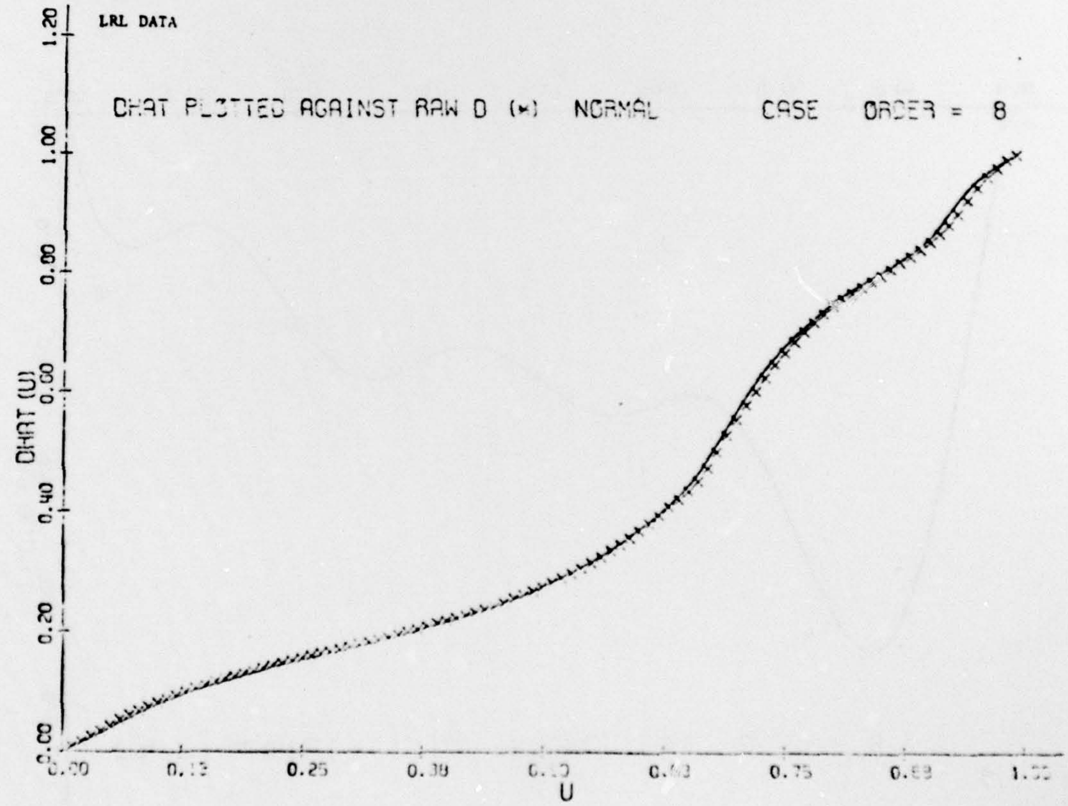


Figure L

LRL DATA 285-555

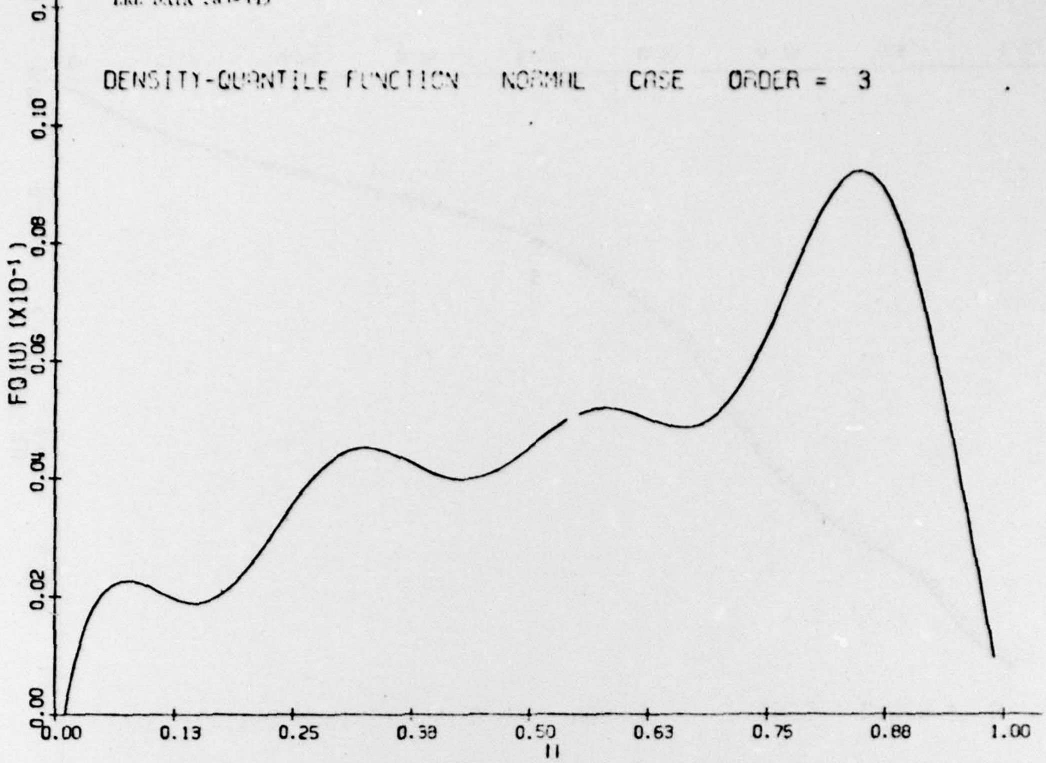
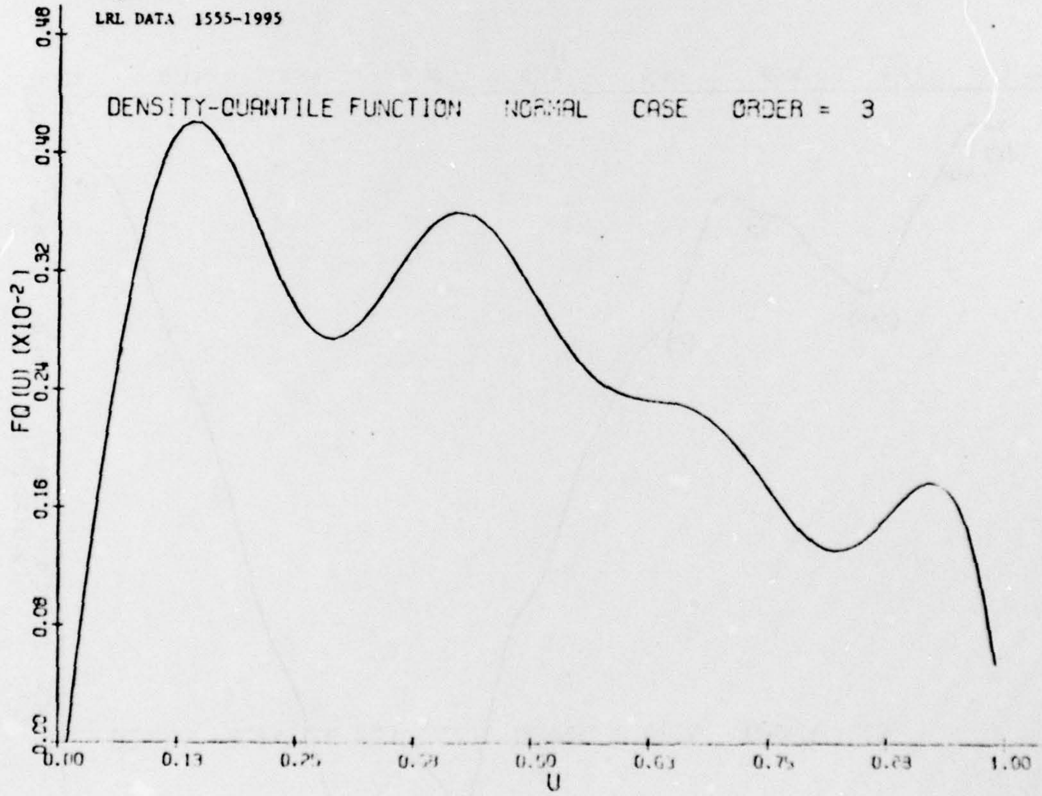
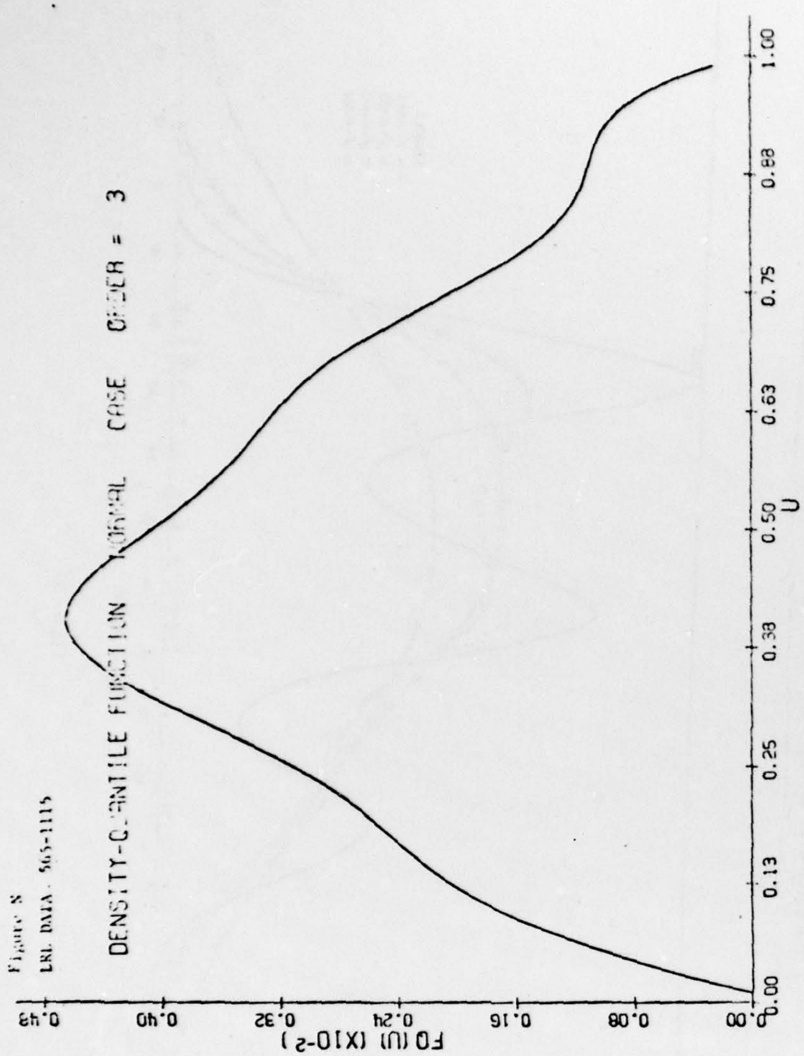


Figure M

LRL DATA 1555-1995





Appendix

Figures 1-3 (with their captions) and Tables 1 and 2 reproduced from the manuscript by Good and Gaskins.

Captions for Figures.

Figure 1. The LRL data, the fitted density of $f(x)$ if $\delta^* = 0.225$, and the thirteen bumps in $f(x)$. The observed bin frequencies are represented by small circles, but, to avoid cluttering the diagram, some of the circles have been omitted when they would be indistinguishable by eye from the fitted curve. Corresponding to each bump there is a pair of brackets that lie close to and between the corresponding pair of points of inflection. Each bracket is within 10 MeV of a point of inflection.

Figure 2. The best fit to the chondrite data ($\delta^* = 0.030$). The scale of the x axis is that used by Leonard (1978) and is $(y - 20)/16$, where y is the percentage of silica given in Table 2. (Thus $-1.25 < x < 5$ by definition.) The 22 observations are marked by crosses above the x axis. See also Table 5.

Figure 3. Surgery for bumps ix and x of the LRL data with $\delta^* = 0.225$. The brackets are placed approximately at the relevant points of inflection. The results of the two surgeries are shown by the curves through the small triangles and squares.

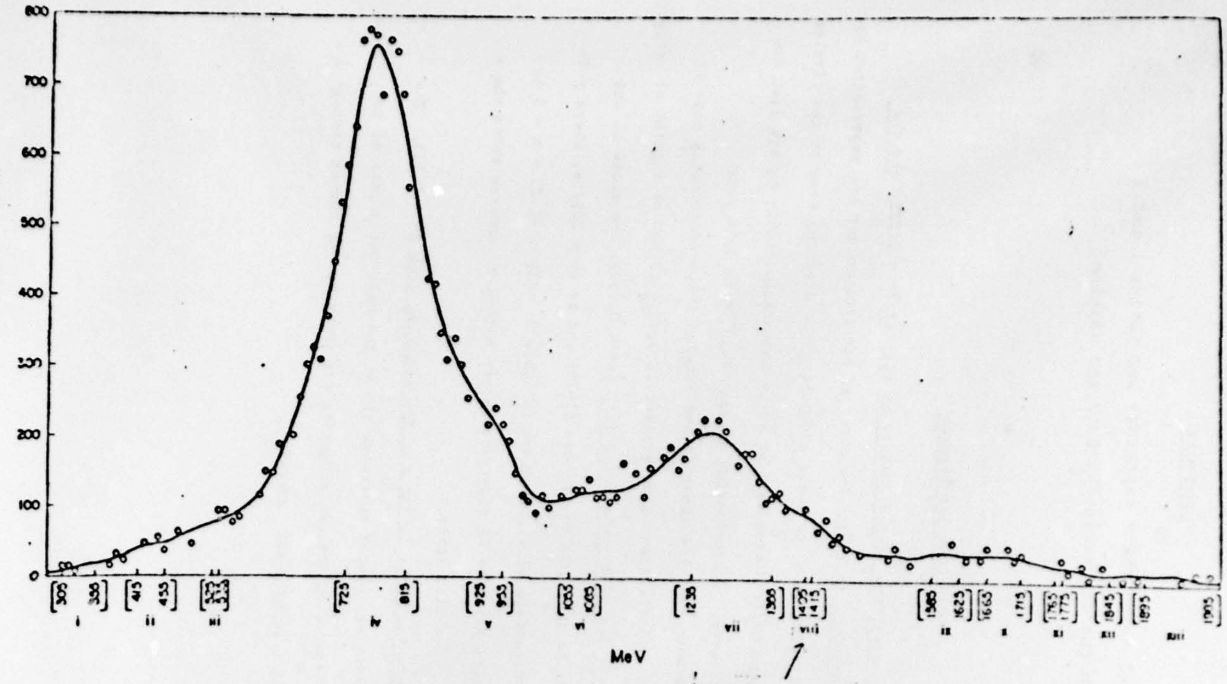


Figure 1.

Glossy photographs of the Figures will be supplied after the paper is accepted for publication.

This figure will be replaced by one with only one curve drawn, namely for $\beta^* = 0.030$. The curve will strongly resemble the one here for $\beta^* = 0.025$.

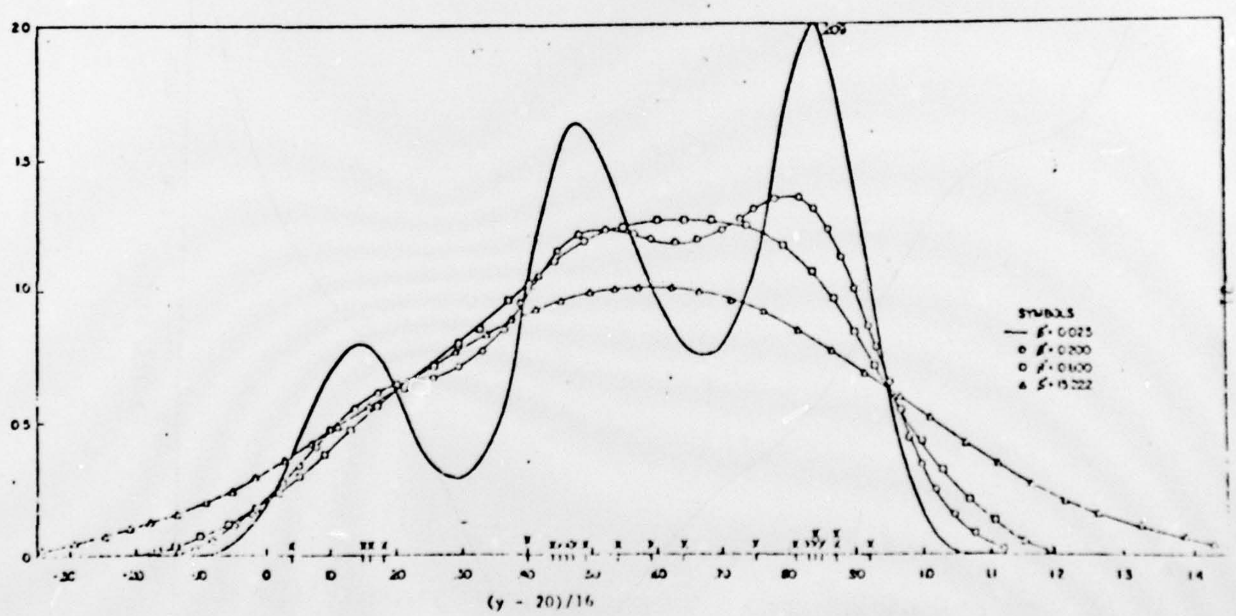


Figure 2.

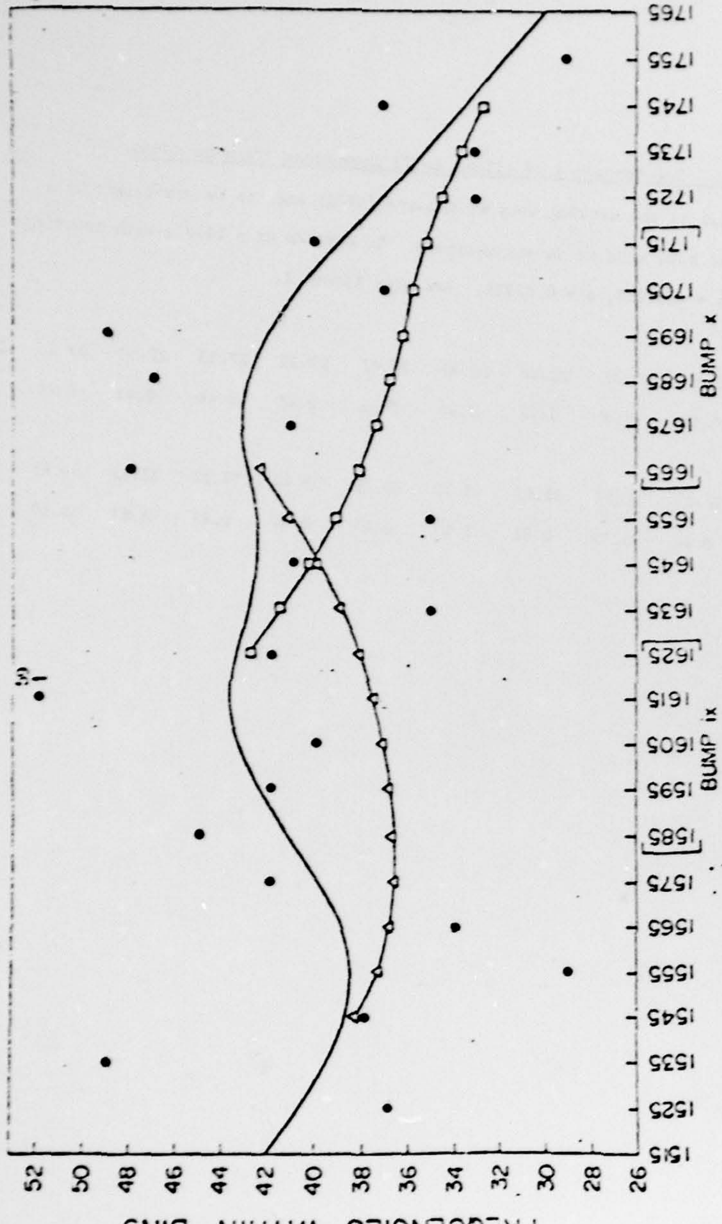


Figure 3.

Table 1. The LRL data of N = 25,752 "events", and the penalized likelihood fit with $\beta^* = 0.225$. Each bin is of width 10 MeV. No observations were made outside the 172 bins shown. Row (a) gives the centers of the bins in MeV; row (b) the observed frequencies; and row (c) the fitted frequencies to the nearest integer. The bumps i to xiii are indicated by bracketed intervals.

(a)	285	295	i [305	315	325	335	345	355]	365	375	385	395	405
(b)	5	11	17	21	15	17	23	25	30	22	36	29	33
(c)	5	8	12	15	18	20	23	25	27	29	32	35	38
	ii [415	425	435	445	455]	465	475	485	495	505	515	iii [525	535]
	43	54	55	59	44	58	66	59	55	67	75	62	98
	42	46	50	53	55	57	60	62	65	69	74	79	84
	545	555	565	575	585	595	605	615	625	635	645	655	665
	94	85	92	102	113	122	153	155	193	197	207	258	305
	88	93	99	106	116	128	143	160	179	200	225	253	285
	675	685	695	705	715	iv [725	735	745	755	765	775	755	795
	332	318	378	457	540	592	646	773	787	783	695	774	759
	321	364	413	471	533	598	658	709	744	761	759	742	709
	805	815]	825	835	845	855	865	875	885	895	905	915	v [925
	692	559	557	499	431	421	353	315	343	306	262	26[illegible]	254
	664	611	557	504	455	412	376	346	321	300	282	26[illegible]	251
	935	945	955]	965	975	985	995	1005	1015	1025	1035	1045	vi [1055
	225	246	225	196	150	118	114	99	121	106	112	122	120
	235	218	199	178	158	140	126	118	113	112	113	115	118
	1065	1075	1085]	1095	1105	1115	1125	1135	1145	1155	1165	1175	1185
	126	126	141	122	122	115	119	166	135	154	120	162	156
	121	124	125	126	127	129	132	136	140	144	148	154	161

1195	1205	1215	1225	vii [1235	1245	1255	1265	1275	1285	1295	1305	1315
175	193	162	178	201	214	230	216	229	214	197	170	181
169	176	184	193	202	210	215	217	214	208	199	188	175
1325]	1335	1345	1355	1365	1375	1385	1395	viii [1405	1415]	1425	1435	1445
133	144	114	120	132	109	108	97	102	89	71	92	58
163	150	138	128	120	113	107	100	94	88	81	74	68
1455	1465	1475	1485	1495	1505	1515	1525	1535	1545	1555	1565	1575
65	55	53	40	42	46	47	37	49	38	29	34	42
62	56	52	48	45	43	42	41	40	39	38	39	40
ix [1585	1595	1605	1615	1625]	1635	1645	1655	x [1665	1675	1685	1695	1705
45	42	40	59	42	35	41	35	48	41	47	49	37
41	42	43	44	43	43	43	43	43	43	43	42	41
1715]	1725	1735	1745	1755	xi [1765	1775]	1785	1795	1805	1815	1825	1835
40	33	33	37	29	26	38	22	27	27	13	18	25
39	37	35	33	32	30	28	27	25	23	22	21	21
xii [1845]	1855	1865	1875	1885	xiii [1895	1905	1915	1925	1935	1945	1955	1965
24	21	16	24	14	23	21	17	17	21	10	14	18
21	20	20	20	20	20	19	19	18	18	17	16	15
1975	1985]	1995										
16	21	6										
13	9	6										

Table 2
is cited
on p.3

Table 2. Percentages y of silica in 22 Chondrites (Ahrens, 1965).

Row (ii) is the scaling used by Leonard (1978) and, to be consistent with him, we have used it in our analysis. It denotes (y - 20)/16 with rounding. Here $\bar{x} = 0.57227$, $s = 0.27226$. See also Figure 2.

y	20.77	22.56	22.71	22.99	26.39	27.08	27.32	27.33	22.57	27.81	28.69
Scaled	0.04	0.15	0.16	0.18	0.40	0.44	0.45	0.46	0.47	0.49	0.54
y	29.36	30.25	31.89	32.88	33.23	33.28	33.40	33.52	33.83	33.95	34.32
Scaled	0.59	0.64	0.75	0.81	0.83	0.84	0.84	0.85	0.87	0.87	0.92